PDA-OTA: Privacy-preserving Distributed Algorithms Over the Air, an OHDSI journey

Authors: Yong Chen[1], Jiayi Tong[1], Chongliang Luo[2], Lu Li[1], Yiwen Lu[1], Hai-Shuo Shu[1]

1. **Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA**
2. **Division of Public Health Sciences, Department of Surgery, Washington University in St. Louis, St. Louis, MO**

Background

Motivated by OHDSI as the next generation open science research consortium for evidence generation and evidence synthesis using distributed real-world data, our team has spent the last six years creating a suite of communication-efficient and heterogeneity-aware distributed algorithms. These are specifically tailored to harness the power of federated real-world data for evidence generation. In addition to these, we have developed user-centric software, and a secure web-based data sharing infrastructure that is designed to meet the unique needs of OHDSI users.

We are proud to announce that we will formally release our software package - PDA[1,2] at this year's OHDSI annual symposium. This includes its user-friendly communication system for distributed learning - PDA-OTA[3,4]. Our innovation is designed to streamline the use of data and foster seamless interaction between users and the system, enabling them to collaborate more efficiently and effectively in their research endeavors.

Methods

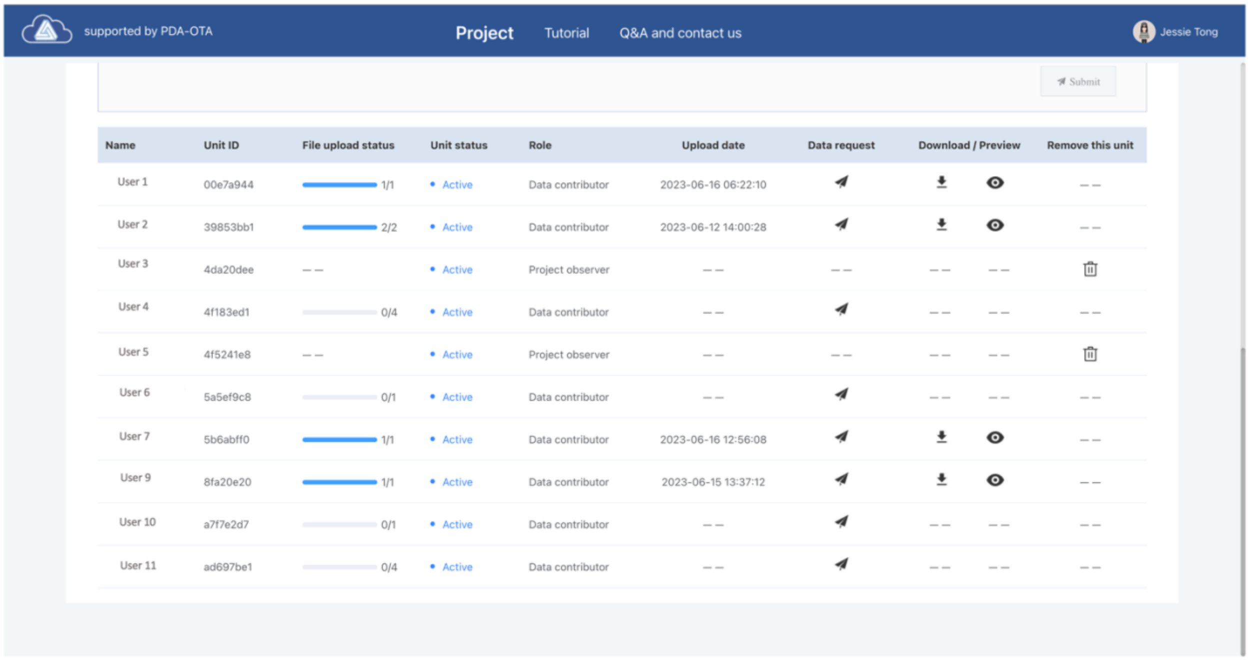
The PDA framework

The PDA framework addresses key challenges in integrating clinical evidence from diverse real-world data sources, such as EHR and claims data. It prioritizes our five key principles: privacy-preservation, communication efficiency, statistical accuracy, heterogeneity-awareness, and implementation readiness. With its versatile applications in association studies, predictive modeling, causal inference, subtyping, and more, PDA has developed over 25 algorithms and has been applied to numerous successful studies. These advancements have significantly contributed to answering critical questions in fields such as pharmacoepidemiology, drug safety studies, clinical decision-making, health policy, and health disparity research.

PDA algorithms have been successfully applied to a wide range of important studies, including investigations into long COVID among children and adolescents, trial emulations for COVID-19 vaccine and booster effectiveness in the pediatric population, racial disparities in Post-acute Sequelae of SARS-CoV-2 Infection (PASC) among children, risk factors for stillbirth[5–7], opioid use disorder (OUD)[8–11], pediatric avoidable hospitalization[12], serious adverse events of colorectal cancer[12], trajectories of Alzheimer's disease (AD)[13], hospitalization and mortality of COVID patients[12,14–16] , risk factors for acute myocardial infarction (AMI)[17], and kidney graft failure[18,19]. By providing this robust suite of tools, PDA aims to streamline the process of evidence generation from distributed real-world data, offering both scientific rigor and simplicity in implementations.

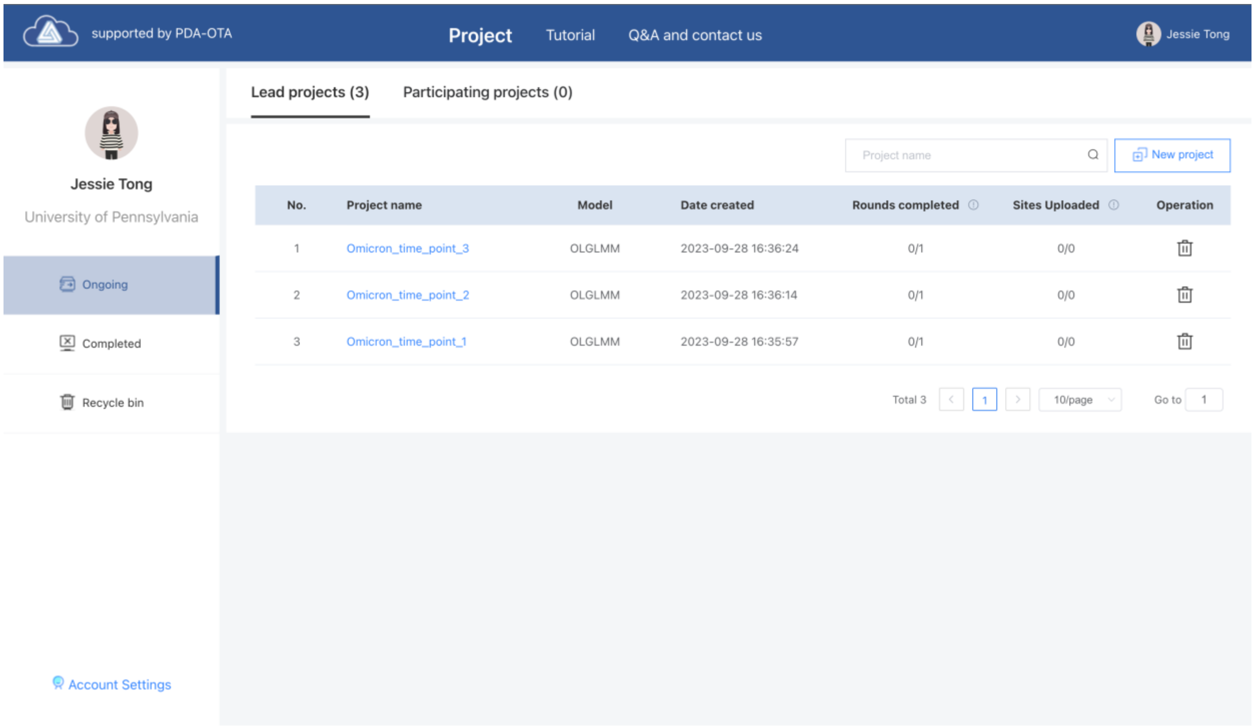
The PDA-OTA: A Tidy Automated Online Workflow with Secure Sharing of Aggregated Data

The PDA-OTA[3] (PDA over the air) is a web-based software developed to support collaborative studies and implement the algorithms within the PDA framework. It enables secure sharing of aggregated data for multi-site studies, ensuring privacy preservation through distributed algorithms. PDA-OTA serves as a unified platform for national and international collaborations, synchronizing project status, offering cloud-based SFTP for data sharing, and generating model-specific tasks for streamlined implementation. Designed with a user-centered approach, it caters to both project leads and participants, allowing them to invite collaborating sites, upload aggregated data, track project status, receive automated email notifications, and generate project summaries automatically. The PDA-OTA provides a tidy and efficient online workflow for secure data sharing and collaboration.



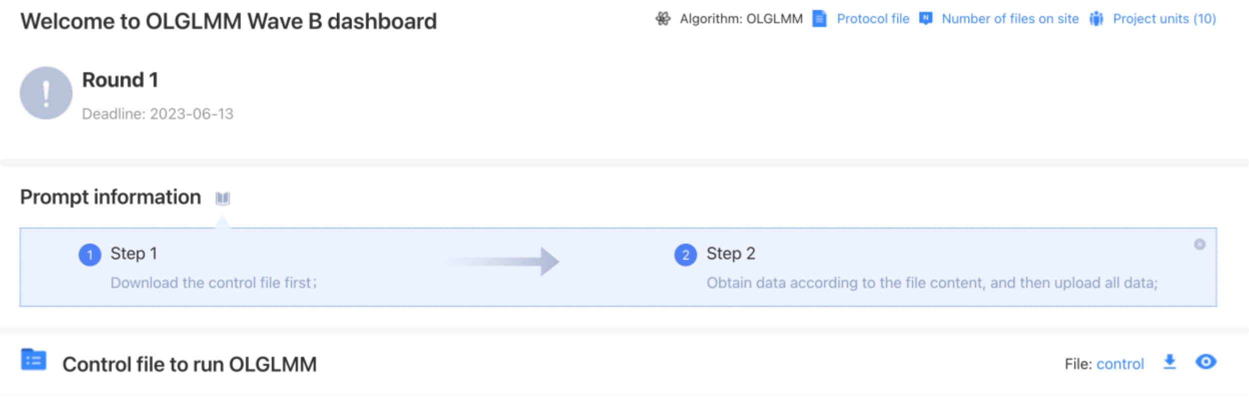
Name	Unit ID	File upload status	Unit status	Role	Upload date	Data request	Download / Preview	Remove this unit
User 1	00e7a944	1/1	• Active	Data contributor	2023-06-16 06:22:10	✈	⬇ 👁	– –
User 2	39853bb1	2/2	• Active	Data contributor	2023-06-12 14:00:28	✈	⬇ 👁	– –
User 3	4da20dee	– –	• Active	Project observer	– –	– –	– – – –	🗑
User 4	4f183ed1	0/4	• Active	Data contributor	– –	✈	– – – –	– –
User 5	4f5241e8	– –	• Active	Project observer	– –	– –	– – – –	🗑
User 6	5a5ef9c8	0/1	• Active	Data contributor	– –	✈	– – – –	– –
User 7	5b6abff0	1/1	• Active	Data contributor	2023-06-16 12:56:08	✈	⬇ 👁	– –
User 9	8fa20e20	1/1	• Active	Data contributor	2023-06-15 13:37:12	✈	⬇ 👁	– –
User 10	a7f7e2d7	0/1	• Active	Data contributor	– –	✈	– – – –	– –
User 11	ad697be1	0/4	• Active	Data contributor	– –	✈	– – – –	– –

Figure 1. PDA-OTA platform



No.	Project name	Model	Date created	Rounds completed	Sites Uploaded	Operation
1	Omicron_time_point_3	OLGLMM	2023-09-28 16:36:24	0/1	0/0	🗑
2	Omicron_time_point_2	OLGLMM	2023-09-28 16:36:14	0/1	0/0	🗑
3	Omicron_time_point_1	OLGLMM	2023-09-28 16:35:57	0/1	0/0	🗑

Figure 2. Projects



Welcome to OLGLMM Wave B dashboard

Round 1

Deadline: 2023-06-13

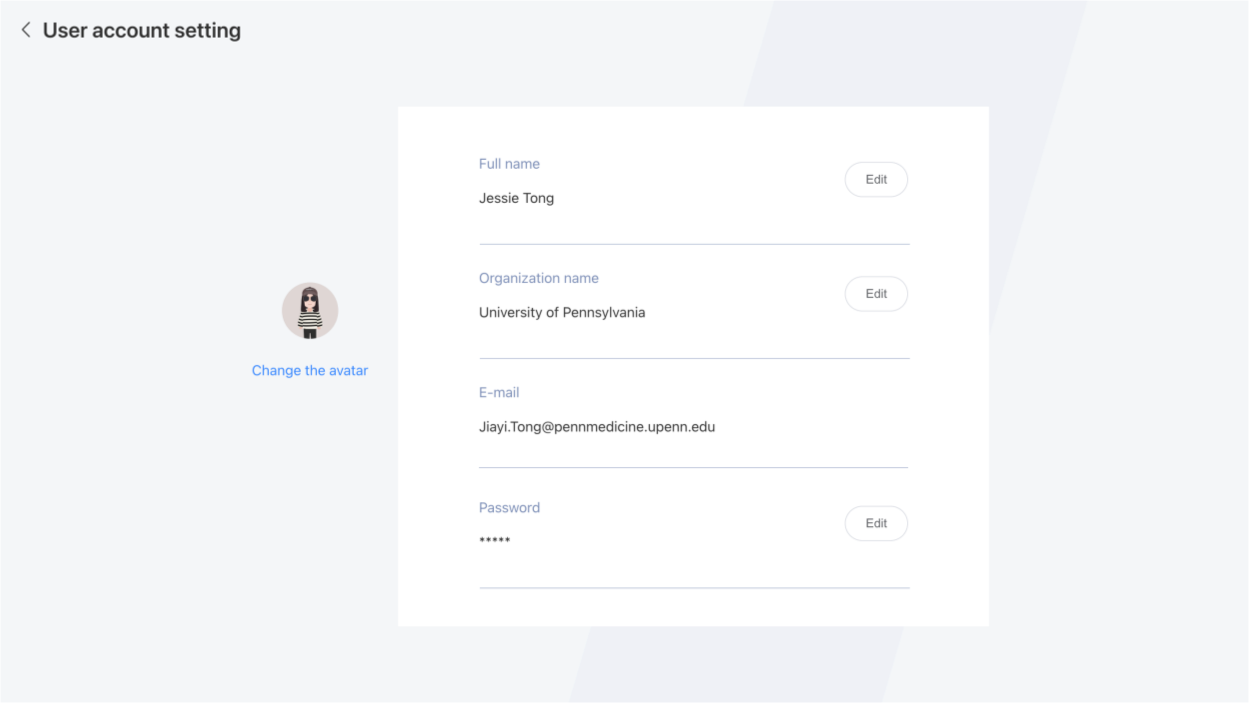
Prompt information

1. Step 1
   Download the control file first;
2. Step 2
   Obtain data according to the file content, and then upload all data;

Control file to run OLGLMM

File: control

Figure 3. Project Initiation



< User account setting

Change the avatar

Full name

Jessie Tong

Edit

Organization name

University of Pennsylvania

Edit

E-mail

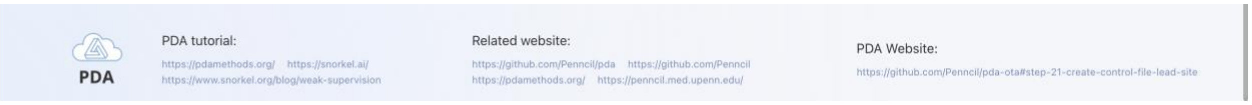
Jiayi.Tong@pennmedicine.upenn.edu

Password

*****

Edit

Figure 4. User account setup page



PDA

PDA tutorial:

https://pdamethods.org/ https://snorkel.ai/
https://www.snorkel.org/blog/weak-supervision

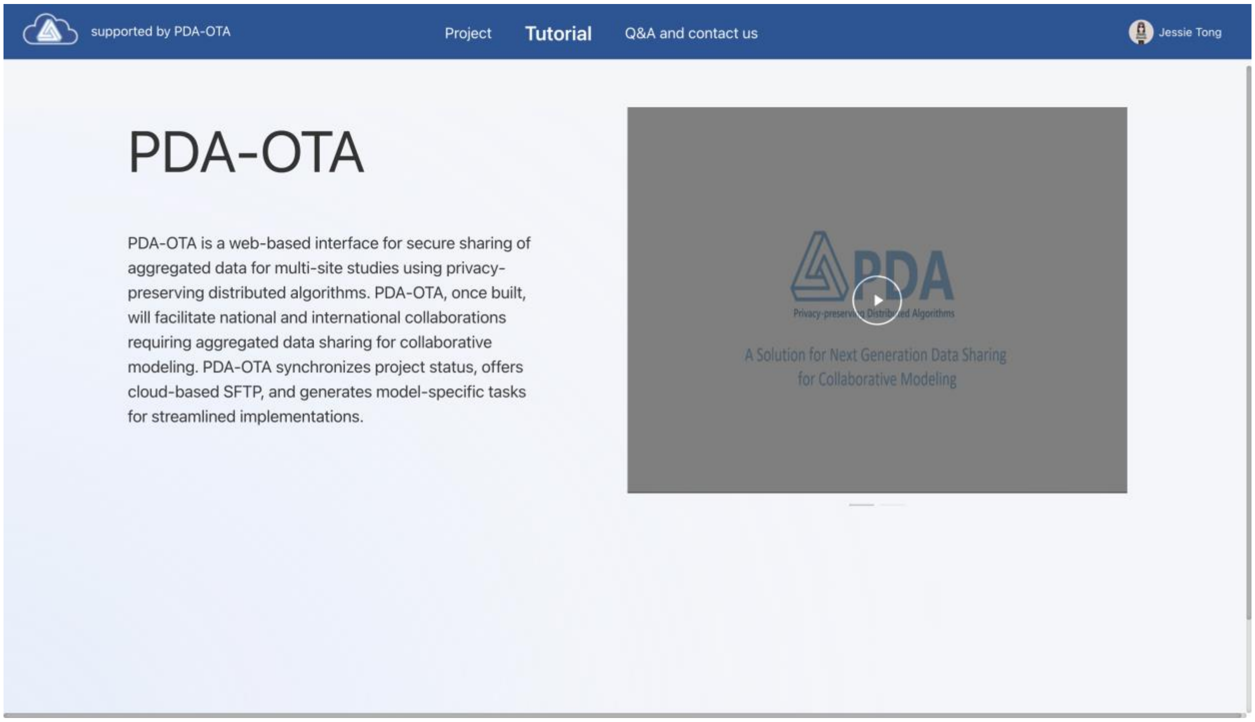
Related website:

https://github.com/Penncil/pda https://github.com/Penncil
https://pdamethods.org/ https://penncil.med.upenn.edu/

PDA Website:

https://github.com/Penncil/pda-ota#step-21-create-control-file-lead-site

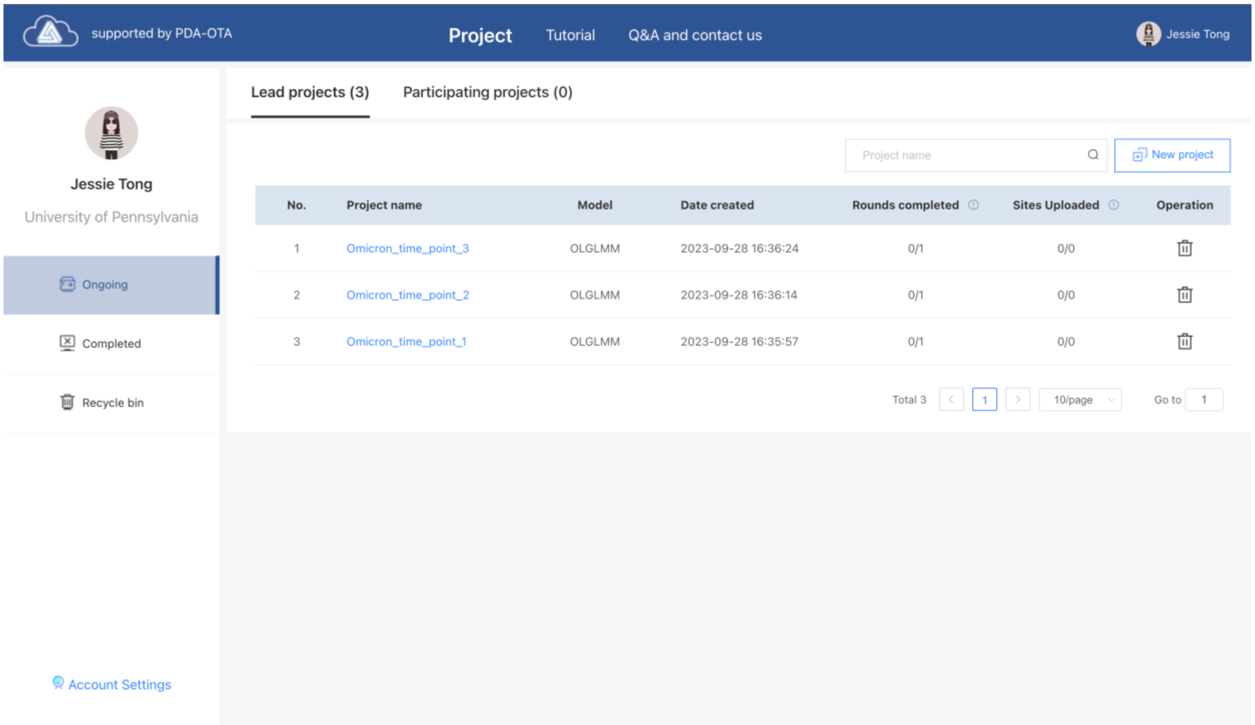
Figure 5. Resources


supported by PDA-OTA
Project
Tutorial
Q&A and contact us
Jessie Tong
PDA
Privacy-preserving Distributed Algorithms
A Solution for Next Generation Data Sharing
for Collaborative Modeling

PDA-OTA

PDA-OTA is a web-based interface for secure sharing of aggregated data for multi-site studies using privacy-preserving distributed algorithms. PDA-OTA, once built, will facilitate national and international collaborations requiring aggregated data sharing for collaborative modeling. PDA-OTA synchronizes project status, offers cloud-based SFTP, and generates model-specific tasks for streamlined implementations.

Figure 6. Tutorial page & video



No.	Project name	Model	Date created	Rounds completed	Sites Uploaded	Operation
1	Omicron_time_point_3	OLGLMM	2023-09-28 16:36:24	0/1	0/0	🗑
2	Omicron_time_point_2	OLGLMM	2023-09-28 16:36:14	0/1	0/0	🗑
3	Omicron_time_point_1	OLGLMM	2023-09-28 16:35:57	0/1	0/0	🗑

Figure 7. Tutorial page & video

Results

PDA and PDA-OTA have revolutionized secure and efficient data sharing in real-world setting of distributed research networks, enabling numerous national and international multi-institutional collaborations. Notably, PDA offers cutting-edge tools such as the Federated Hospital Comparer framework, which employs counterfactual modeling to facilitate accurate and equitable comparisons of hospital performance across decentralized data sources. Furthermore, PDA encompasses a groundbreaking federated learning approach that investigates racial disparities in clinical outcomes linked to differential healthcare access. For instance, our analysis of COVID-19 outcomes uncovered the association between increased mortality among Black patients and the hospitals where they disproportionately receive care. Moreover, our ongoing research on the impact of site of care on kidney transplant outcomes demonstrates the potential to mitigate disparities through a federated learning framework. With a commitment to advancing the missions of OHDSI, we continuously refine and enhance the capabilities of PDA and PDA-OTA as the next generation data science for clinical evidence generation and evidence synthesis.

References/Citations

1. CRAN - Package pda. https://cran.r-project.org/web/packages/pda/index.html.
2. PDA website. https://pdamethods.org/.
3. PDA-OTA. https://pda-ota.pdamethods.org/login.
4. Penncil/pda-ota. https://github.com/Penncil/pda-ota.
5. Tong, J. *et al.* Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data. in *Pacific Symposium on Biocomputing* vol. 25 695–706 (World Scientific Publishing Co. Pte Ltd, 2020).
6. Duan, R., Boland, M. R., Moore, J. H. & Chen, Y. ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. in *Biocomputing 2019* 30–41 (WORLD SCIENTIFIC, 2018). doi:10.1142/9789813279827_0004.
7. Duan, R., Boland, M. R., Moore, J. H. & Chen, Y. ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *Pac. Symp. Biocomput.* **24**, 30–41 (2019).
8. Luo, C. *et al.* ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data. *Sci Rep* **12**, 1–8 (2022).
9. Tong, J. *et al.* Identifying Clinical Risk Factors for Opioid Use Disorder using a Distributed Algorithm to Combine Real-World Data from a Large Clinical Data Research Network. *AMIA Annual Symposium Proceedings* **2020**, 1220 (2020).
10. Penncil/ADAP. https://github.com/Penncil/ADAP.
11. Liu, X. *et al.* Multisite learning of high-dimensional heterogeneous data with applications to opioid use disorder study of 15,000 patients across 5 clinical sites. *Sci Rep* **12**, (2022).
12. Edmondson, M. J. *et al.* An efficient and accurate distributed learning algorithm for modeling multi-site zero-inflated count outcomes. *Sci Rep* **11**, 1–17 (2021).
13. Duan, R. *et al.* Leverage Real-world Longitudinal Data in Large Clinical Research Networks for Alzheimer’s Disease and Related Dementia (ADRD). *AMIA Annu Symp Proc* **2020**, 393–401 (2020).
14. Luo, C. *et al.* DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. *Nat Commun* **13**, 1–10 (2022).
15. Luo, C. *et al.* dPQL: a lossless distributed algorithm for generalized linear mixed model with application to privacy-preserving hospital profiling. *Journal of the American Medical Informatics Association* ocac067 (2022) doi:10.1093/jamia/ocac067.
16. Edmondson, M. J. *et al.* Distributed Quasi-Poisson Regression Algorithm for Modeling Multi-Site Count Outcomes in Distributed Data Networks. *J Biomed Inform* 104097 (2022).
17. Duan, R. *et al.* Learning from local to global-an efficient distributed algorithm for modeling time-to-event

data. *Journal of the American Medical Informatics Association* **27**, 1028–1036 (2020).

18. Penncil/dGEM. https://github.com/Penncil/dGEM.
19. Penncil/dGEM-disparity. https://github.com/Penncil/dGEM-disparity.